

Extensions of stability selection using subsamples of observations and covariates

Andre Beinrucker · Ürün Dogan · Gilles Blanchard

Abstract We introduce extensions of stability selection, a method to stabilise variable selection methods introduced by Meinshausen and Bhlmann (J R Stat Soc 72:417-473, 2010). We propose to apply a base selection method repeatedly to random subsamples of observations and subsets of covariates under scrutiny, and to select covariates based on their selection frequency. We analyse the effects and benefits of these extensions. Our analysis generalizes the theoretical results of Meinshausen and Bhlmann (J R Stat Soc 72:417-473, 2010) from the case of half-samples to subsamples of arbitrary size. We study, in a theoretical manner, the effect of taking random covariate subsets using a simplified score model. Finally we validate these extensions on numerical experiments on both synthetic and real datasets, and compare the obtained results in detail to the original stability selection method.

Keywords variable selection · stability selection · subsampling

A preliminary version of this work was presented at the conference DAGM 2012 (Beinrucker et al., 2012b).

Andre Beinrucker, Gilles Blanchard
University of Potsdam
Am Neuen Palais 10, 14469 Potsdam, Germany
E-mail: {andre.beinrucker,gilles.blanchard}@uni-potsdam.de

Ürün Dogan
Microsoft/Skype Labs
2 Waterhouse Square, 140 Holborn EC1N2ST London
Unided Kingdom
E-mail: urundogan@gmail.com

1 Introduction

1.1 Motivation

In many applications a very large number of covariates are observed, of which only a few carry information about an outcome of interest. Variable selection techniques aim at identifying such relevant covariates (for a review see Guyon, 2006). Usually, variable selection aims at one of two goals: to identify informative covariates in order to get scientific insight into the data and the process that generated the outcome; or to use the covariates identified as relevant in order to predict the outcome. In this work we primarily focus on the identification of informative covariates but also consider prediction results using real data. We consider *variable selection* (also called *feature selection* in computer science-related communities) as a part of the broader field of dimensionality reduction.

Many variable selection methods share the common drawback of being unstable with respect to small changes of the data: if one estimates the set of relevant covariates on different sets of observations coming from the same source, the result can vary significantly. While this is not necessarily of concern if prediction is the goal, it makes the identification of relevant covariates very difficult. One approach to overcome this problem is *stability selection* (Meinshausen and Bühlmann, 2010). It consists in applying repeatedly a variable selection method to randomly chosen subsamples of half size of the observations. The final selection is obtained by picking only those covariates whose selection frequency across repetitions exceeds a certain threshold. This threshold can be chosen such that (under some assumptions) the expected number of false positive selections is guaranteed to be below a chosen value.

1.2 Contributions

In the remainder of the paper we will refer to the variable selection method that is repeatedly applied to data subsamples as the *base method*. Similarly to Meinshausen and Bühlmann (2010), we construct a method that can be applied on top of an arbitrary base method, which is considered as a black box.

We propose to extend the central idea of stability selection in two natural directions. First, Meinshausen and Bühlmann (2010) use random samples containing half of the observations of the full dataset. Instead, we choose some integer $L > 1$ and draw subsamples of size $\frac{1}{L}$ of the full sample size. More precisely, we randomly partition the observations into disjoint subsamples, extending the approach of complementary pairs stability selection - CPSS (Shah and Samworth, 2013). We investigate theoretically the behaviour of the expected number of false positive selections depending on the number of subsamples. In addition, we perform extensive simulation studies to compare the number of correct variables recovered for artificial and semi-synthetic datasets.

Secondly, Meinshausen and Bühlmann (2010) remarked from empirical comparisons that stability selection can be improved by randomising the base method. We propose a randomization by simply applying the base method to random subsets of covariates. We obtain these subsets by randomly partitioning the covariates into disjoint subsets. Because the effect of doing so depends on the base method being used, it is difficult to analyse it theoretically in full generality. In this work, we restrict our theoretical analysis to a simplified toy model, in which we assume that for each covariate there exists a latent score reflecting its informativeness about the outcome of interest. Furthermore, we assume that the base method has access to noisy observations of these scores, and outputs the covariate with the largest observed score. We investigate how the probability of selecting a noninformative covariate (false positive) is influenced by the size of the random subset of covariates used. Besides the theoretical analysis of this toy model, we performed simulation studies similar to the ones in the investigation of the subsampling of observations.

We call the method that combines the two extensions proposed *extended stability selection*. To summarize, it applies the base method repeatedly to randomly chosen subsets of the observations and covariates and finally ranking covariates by their selection frequency. There are two parallel goals for this extension. The first goal is to improve the precision of the selection, that is to reduce the number of false positives. The second goal is to reduce the computational complexity of stability

selection. Indeed, each call of the base method is restricted to a subset of observations *and* covariates; this reduces the memory requirements of the method. If the complexity of the base method grows faster than linearly in the data size (number of observations times number of covariates), the total computation cost is also reduced. This is particularly appealing if the base method needs to load the data that it operates on into memory, which can be infeasible for large data matrices, but easy for smaller submatrices. Furthermore, this naturally allows for parallelization of the method, since these submatrices can be processed independently.

1.3 Overview of results

Concerning the subsampling of observations, we obtain a bound on the expected number of false positives, depending on the size of the subsamples. This bound sharply generalizes Theorem 1 of Meinshausen and Bühlmann (2010) and Theorem 1 of Shah and Samworth (2013). Our results suggest that there is a trade-off between improving the selection of covariates for each individual subsample by using a smaller number of larger subsamples, and improving the final selection by averaging over a larger number of smaller, independent subsamples. This finding is in line with general insights on subsampling methods (Politis et al., 1999) and cross-validation (Arlot and Celisse, 2010, Section 10.3). Even though our empirical comparison shows only small differences, a significant advantage of our proposed subsampling extension is that it has much less computational and memory requirements compared to the original stability selection or CPSS.

For the randomization of the base method obtained by taking disjoint subset of covariates, the theoretical analysis of our simplified score model shows that under certain assumptions, there exists an optimal size for the randomly chosen covariate subsets. Our empirical results support this finding: such randomization generally improves the performance, unless the subset size is too small.

1.4 Organization of the paper

In Section 2 we give a detailed description of the algorithm proposed, including the base methods considered in the experiments. The theoretical analysis is presented in Section 3. We motivate the use of small observation subsamples in Section 3.1 and investigate the randomization of the base method in Section 3.2. Experimental results are given in Section 4, where we measure the performance of the algorithm in selecting in-

formative covariates in Section 4.1 and 4.2 and apply our method in an image classification setting in Section 4.3. We conclude our work in Section 5 with a summary and a discussion.

2 Methods

2.1 Description of the algorithm

In the sequel we assume to be given a dataset \mathcal{D} containing N observations $(X^{(i)}, Y^{(i)})_{i=1, \dots, N}$, each observation consisting of D covariates (X_1, \dots, X_D) and an outcome Y of interest. We choose T , the number of times we repeat the random partitioning of the data and a threshold $\tau \in (0, 1)$ that indicates the fraction of observation subsamples in which a covariate needs to be chosen in order to enter the final selection. The number of observations and covariates that we apply our base method on is determined by the parameters (L, V) . We use random observation subsamples of size $\lfloor \frac{N}{L} \rfloor$ and covariate subsets of size $\lfloor \frac{D}{V} \rfloor$ or $\lfloor \frac{D}{V} \rfloor + 1$.

We denote by $S^{\text{base}}(X_{\mathcal{F}}^{(\mathcal{S})}, Y^{(\mathcal{S})})$ the output of the base method applied to the restriction of the data \mathcal{D} to observations with indices $\mathcal{S} \subset \{1, \dots, N\}$ and covariates with indices $\mathcal{F} \subset \{1, \dots, D\}$. We denote by $\Pi_{L,V}^{\text{SFS}}(d)$ the selection frequency of covariate d where the superscript SFS stands for *stability feature selection*. We give the pseudo-code of the method we propose below. Note that we recover the original stability selection algorithm (more precisely, the CPSS algorithm of Shah and Samworth, 2013) for $L = 2, V = 1$.

Parameters:

- Number of iterations T
- Number of observation subsamples per iteration L
- Number of covariate subsets per iteration V
- Threshold $\tau \in (0, 1)$

Input: $\mathcal{D} = (X^{(i)}, Y^{(i)})_{i=1, \dots, N}$ (with $X^{(i)} \in \mathbb{R}^D$.)

Initialization: selection frequencies $\Pi_{L,V}^{\text{SFS}}(d) = 0$, $d = 1, \dots, D$

for $t = 1$ **to** T **do**

Draw L disjoint random subsamples $\mathcal{S}(1), \dots, \mathcal{S}(L)$ of size $\lfloor \frac{N}{L} \rfloor$ without repetition of $\{1, \dots, N\}$.

Partition $\{1, \dots, D\}$ into V disjoint random subsets $\mathcal{F}(1), \dots, \mathcal{F}(V)$ of size $\lfloor \frac{D}{V} \rfloor$ or $\lfloor \frac{D}{V} \rfloor + 1$.

for $i = 1, \dots, L$; $j = 1, \dots, V$ **do**

$\mathcal{G} := S^{\text{base}}(X_{\mathcal{F}(j)}^{(\mathcal{S}(i))}, Y^{(\mathcal{S}(i))})$

for all $d \in \mathcal{G}$ **do**

$\Pi_{L,V}^{\text{SFS}}(d) \leftarrow \Pi_{L,V}^{\text{SFS}}(d) + 1/LT$

end for

end for

end for

return set of indices in final selection

$$S_{L,V,\tau}^{\text{SFS}} := \{d : \Pi_{L,V}^{\text{SFS}}(d) \geq \tau\}.$$

2.2 Comparison to previous work

Statistical methods can be applied to subsamples of the data in various ways. A classical way in this context is the bootstrap (Efron, 1979), where subsamples are drawn with replacement. In contrast, stability selection and our extension follow the idea of subsampling without replacement (Politis et al., 1999) and are strongly related to subagging (Bühlmann and Yu, 2002).

Several approaches have been developed to combine variable selection and subsampling of observations. Sauerbrei and Schumacher (1992) investigated bootstrapping variable selection methods in the Cox regression model. Further, several methods have been proposed where a predictor that does variable selection intrinsically is applied to subsamples of the data. The selection obtained is then used as input to the final prediction method. Bi et al. (2003) used linear a support vector machines (SVM) to select variables and train a kernel SVM on them. Similarly, in Bolasso (Bach, 2008), the Lasso method is applied to bootstrapped subsamples of the data. The final selection is obtained as intersection of the sets of selected covariates on the bootstraps. On this final selection, Lasso is applied again to obtain a final estimate of the coefficients. Similarly, in random Lasso (Wang et al., 2011) the importance of covariates is estimated by applying Lasso to bootstrap subsamples of the data. Afterwards Lasso is applied to a random selection of the covariates, the probability to be included is proportional to the measure obtained in the first step.

Meinshausen and Bühlmann (2010) proposed stability selection, a method that combines subsampling with variable selection. In their analysis the authors give a bound for the number of false positives under some simplifying assumptions. Their work was the basis for numerous follow-up studies by theoreticians and practitioners.

Shah and Samworth (2013) introduced complementary pairs stability selection (CPSS), a variant of stability selection which uses not only subsamples of size $\lfloor \frac{N}{2} \rfloor$ of the data, but also their complements. They loosen the assumptions of the original method and give bounds on both errors of the selection procedure, false positives and false negatives. Our work is an extension of their results, as we apply our base method to complementary subsamples as well.

Stability selection has been applied in various disciplines. Fields of application include genome-wide association studies (Alexander and Lange, 2011; He and

Lin, 2011), biomarker discovery (He and Yu, 2010) and inference of gene regulatory networks (Haury et al., 2012).

The idea of repeatedly applying a statistical method to covariate subsets has been investigated before. One famous example is Random Forest (Breiman, 2001) where decision trees are build on covariate subsets. Each decision tree can be regarded as a variable selection method as well. Recently, Hinton et al. (2012) remarked that omitting randomly chosen covariates in the training of a neuronal network improves classification accuracy on test data drastically. A different approach to reduce the dimensionality of the problem is to cluster the covariates and apply a variable selection method to cluster representatives (Bühlmann et al., 2013).

2.3 Base methods

In stability selection any variable selection method can in principle be used as a base method. In this section, we describe variable selection methods in general and give details about two selection methods that we used as a base method for our extended stability selection experiments. Base methods are described below for the full sample \mathcal{D} for simplicity.

Variable selection methods can be classified into filters, wrapper and embedded methods (Guyon, 2006). Computationally most efficient are usually filter methods, which perform variable selection independently of the specific statistical treatment that might be applied afterwards. Examples are methods based on simple univariate correlation between covariates and the outcome of interest or mutual information (see, e.g., Cover and Thomas, 2006, Chapter 2). Wrapper methods evaluate the relevance of a subset of covariates using the output of the ensuing statistical treatment (typically regression or discrimination) computed on the subset of covariates only. Embedded methods perform variable selection and prediction simultaneously.

To assess the performance of the proposed methods, we choose two popular methods as base method: CMIM - conditional mutual information maximisation (Fleuret, 2004) from the class of filters and Lasso (Tibshirani, 1996) from the class of embedded methods. Lasso was also used as a base method in the original work of Meinshausen and Bühlmann (2010)

2.3.1 CMIM - conditional mutual information maximization

Intuitively, the aim of the CMIM algorithm (Fleuret, 2004) is to find a subset of covariates of given size K that maximizes the amount of information that the

selected covariates $X_{\nu(1)}, \dots, X_{\nu(K)}$ contain about the outcome Y . It usually finds a good trade-off between relevance and redundancy of the selected covariates and is much faster than many competing mutual information based variable selection methods.

To make the computation feasible, the CMIM algorithm does not look directly for the set of covariates that globally maximizes the mutual information with the target, but performs greedy selection by iteratively selecting the covariate that has the largest mutual information with the target, conditional to the set of already selected covariates. Furthermore, the latter quantity is approximated by a simpler upper bound, namely, the minimal information about the target that a candidate covariate adds to any of the already selected covariates, taken individually (rather than jointly). The final algorithm takes the following simple form:

$$\nu(1) = \underset{d \in \{1, \dots, D\}}{\text{Arg Max}} \hat{I}(X_d; Y); \quad (1)$$

$$\nu(\ell) = \underset{d \in \{1, \dots, D\}}{\text{Arg Max}} \min_{j < \ell} \hat{I}(X_d; Y | X_{\nu(j)}), \quad \ell > 1, \quad (2)$$

where $\hat{I}(X; Y)$ denotes an estimator of the mutual information of X and Y and $\hat{I}(X; Y | Z)$ an estimator of the mutual information between X and Y given Z . To speed up the computation, one can use a fast implementation of the algorithm (Fleuret, 2004).

2.3.2 Lasso

The Lasso problem (Tibshirani, 1996) is to find an ℓ_1 regularized solution for the least squares problem in the linear model. It can be stated as

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^D}{\text{Arg Min}} \sum_{i=1}^N (Y^{(i)} - \langle X^{(i)}, \beta \rangle)^2 + \lambda |\beta|, \quad (3)$$

where $\lambda \in \mathbb{R}_{>0}$ is some regularization parameter.

Thanks to the geometric properties of the ℓ_1 -Norm, solutions of the Lasso problem tend to be sparse, meaning that only a few coefficients of β are non-zero. This property allows us to use Lasso solvers as variable selection methods, and to define for any $\lambda > 0$ the set of selected covariates as the ones corresponding to non-zero coefficients. Additionally to the selection of relevant covariates, a solution $\hat{\beta}_\lambda$ of the Lasso problem can be used to predict an outcome for new data. Examples for popular Lasso solvers are LARS (Hastie and Efron, 2012) and GLMNET (Friedman et al., 2010).

3 Analysis

In this section, we analyse the effect of the size of the observations subsamples as well as the effect of taking

random covariate subsets on the performance of the final method.

3.1 Subsampling of observations

In this section the N observations of the full sample \mathcal{D} are always assumed to be drawn i.i.d. from an underlying, unknown generating distribution. As we do not take covariate subsets in this section, we fix $V = 1$ and suppress the dependence on V in the notation. We recall that we repeat T times the random draw of L disjoint observation subsamples of equal size without repetition; we denote $\mathcal{S}(\ell, t)$ the ℓ -th subsample of observations indices in the t -th drawing. On the $L \times T$ subsamples that we obtain, the selection frequency of covariate d is

$$\Pi_L^{SFS}(d) := \frac{1}{TL} \sum_{t=1}^T \sum_{\ell=1}^L \mathbf{1}\{d \in S^{\text{base}}(X^{(\mathcal{S}(\ell, t))}, Y^{(\mathcal{S}(\ell, t))})\}.$$

Thresholding this quantity, we obtain the output of the stability selection procedure. For any $\tau \in (0, 1)$ we define

$$S_{L, \tau}^{SFS} := \{d : \Pi_L^{SFS}(d) \geq \tau\}.$$

To evaluate the performance of the method, we need to define the set of covariates that we would like to be excluded from our selection, i.e. the covariates that we consider false positives if they are selected. Since the base method is treated here as a black box and otherwise unspecified, we have to trust that, *on average* (over a random i.i.d. sample), the base method selects relevant covariates more frequently than irrelevant ones. Consider virtually drawing an independent, i.i.d. set of observations of size $\lfloor \frac{N}{L} \rfloor$ and denote the random output of the base method on this sample by S_L^{base} . Then we define for each covariate d

$$p_L(d) := \mathbb{P}[d \in S_L^{\text{base}}],$$

its probability to be selected by the base method using a sample of size $\lfloor \frac{N}{L} \rfloor$.

The quantities $p_L(d)$ give us a yardstick to rank the covariates by relevance; accordingly, for any given threshold $\theta \in (0, 1)$ we define

$$A_{\theta, L} := \{d : p_L(d) \leq \theta\}$$

the set of uninformative covariates at base selection probability lower than θ . In this definition, and the assumption that the probability of selection under the base method reflects the true relevance, we follow the general approach of Shah and Samworth (2013); the relation to the assumptions of Meinshausen and Bühlmann

(2010) is discussed below. Observe that, since each subsample appearing in the definition $\Pi_L^{SFS}(d)$ is individually an i.i.d. sample of size $\lfloor \frac{N}{L} \rfloor$, it follows that $\Pi_L^{SFS}(d)$ is an unbiased estimate of $p_L(d)$.

The following theorem bounds in Equation (4) the ratio of the expected number of false positive selections of $S_{L, \tau}^{SFS}$ compared to the total number of uninformative covariates (false positive rate), as well as compared to the expected number of false positive selections of the base method applied on a single subsample in Equation (5). A corresponding result for the false negatives is available in Equations (6) and (7).

Theorem 1 *Let $L \in \mathbb{N}$ and $\tau \in (0, 1)$. Denote $\forall p, q \in (0, 1)$ $D(p, q) := p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$ the Kullback-Leibler divergence between two Bernoulli variables of parameters p and q . For two integers $a \leq b$ denote $\{a..b\}$ the integer interval with endpoints a, b . Depending on the choice of θ and τ we have two cases with two results each:*

If $\theta < \tau$ we have

$$\frac{\mathbb{E}[|S_{L, \tau}^{SFS} \cap A_{\theta, L}|]}{|A_{\theta, L}|} \leq \min_{\ell_0 \in \{\lfloor L\theta \rfloor .. \lfloor L\tau \rfloor\}} \left(\frac{L - \ell_0 + 1}{\tau L - \ell_0 + 1} \right) \exp \left(-LD \left(\frac{\ell_0}{L}, \theta \right) \right) \quad (4)$$

and

$$\frac{\mathbb{E}[|S_{L, \tau}^{SFS} \cap A_{\theta, L}|]}{\mathbb{E}[|S_L^{\text{base}} \cap A_{\theta, L}|]} \leq \min_{\ell_0 \in \{\lfloor L\theta \rfloor + 1 .. \lfloor L\tau \rfloor\}} \left(\frac{L - \ell_0 + 1}{\tau L - \ell_0 + 1} \right) \frac{\exp(-LD(\frac{\ell_0}{L}, \theta))}{\theta}, \quad (5)$$

Similarly, if $\tau < \theta$ we have

$$\frac{\mathbb{E}[|(S_{L, \tau}^{SFS})^c \cap (A_{\theta, L})^c|]}{|(A_{\theta, L})^c|} \leq \min_{\ell_0 \in \{\lfloor L\tau \rfloor .. \lfloor L\theta \rfloor\}} \left(\frac{\ell_0 + 1}{\ell_0 - \tau L + 1} \right) \exp \left(-LD \left(\frac{\ell_0}{L}, \theta \right) \right), \quad (6)$$

and

$$\frac{\mathbb{E}[|(S_{L, \tau}^{SFS})^c \cap (A_{\theta, L})^c|]}{\mathbb{E}[|(S_L^{\text{base}})^c \cap (A_{\theta, L})^c|]} \leq \min_{\ell_0 \in \{\lfloor L\tau \rfloor .. \lfloor L\theta \rfloor - 1\}} \left(\frac{\ell_0 + 1}{\ell_0 - \tau L + 1} \right) \frac{\exp(-LD(\frac{\ell_0}{L}, \theta))}{1 - \theta}, \quad (7)$$

where A^c denotes the complement of a set A (in $\{1, \dots, D\}$).

For the special case $L = 2$ we recover the results of Shah and Samworth (2013, Theorem 1) by choosing $\ell_0 = 2$ in Equation (4) and $\ell_0 = 0$ in (6). In the following corollary we formulate our results under the assumptions and in the notation of Meinshausen and Bühlmann (2010, Theorem 1):

Corollary 1 Suppose we are given a set of noise covariates \mathcal{N} and a set of signal covariates. Assume that all noise covariates have the same probability to be selected by the base method. Assume further that the base variable selection method has a larger probability to select any informative covariate than random guessing. We denote $q_L = \mathbb{E}[|S_L^{\text{base}}|]$. Then for any $\tau > \frac{q_L}{D}$:

$$\frac{\mathbb{E}[|S_{L,\tau}^{SFS} \cap \mathcal{N}|]}{|\mathcal{N}|} \leq \min_{\ell_0 \in \{\lceil \frac{Lq_L}{D} \rceil, \dots, \lceil \tau L \rceil\}} \left(\frac{L - \ell_0 + 1}{\tau L - \ell_0 + 1} \right) \exp \left(-LD \left(\frac{\ell_0}{L}, \frac{q_L}{D} \right) \right) \quad (8)$$

If we choose $\ell_0 = L = 2$ and we use that $\exp(-2D(1, \theta)) = \theta^2$ we recover the order $O(q_L^2/D)$ of the bound of Meinshausen and Bühlmann (2010, Theorem 1) as well as the constraint $\tau > 1/2$ there.

Bound (8) involves a minimum over the allowed values of ℓ_0 , which is merely a technical parameter in the bound. In order to make the bound more readable, we can pick a specific value of ℓ_0 as follows. As the exponential term in the bound is monotonically decreasing in ℓ_0 , the largest allowed value $\ell_0 = \lceil \tau L \rceil$ seems a natural choice. However, one should ensure that the multiplicative term in front does not become too large. For this, choose τ and L such that τL is an integer; then we have $\frac{\ell_0 - 1}{L} = \frac{\tau L - 1}{L} = \tau - \frac{1}{L}$. In this case, Corollary 1 simplifies as follows.

Corollary 2 Under the assumptions of Corollary 1, if τL is an integer and if we choose $\ell_0 = \tau L$, then we have as a direct consequence of Corollary 1 for any $\tau > \frac{q_L}{D}$:

$$\frac{\mathbb{E}[|S_{L,\tau}^{SFS} \cap \mathcal{N}|]}{|\mathcal{N}|} \leq (L(1 - \tau) + 1) \exp \left(-LD \left(\tau, \frac{q_L}{D} \right) \right). \quad (9)$$

Remark 1. As the expected number of false positive selections decays exponentially with L , it is tempting to conclude that L should be chosen as large as possible. But one should not forget that the parameter L has an important influence on base selection probabilities $p_L(d)$ as well. This influence strongly depends on the particular base method used, and we cannot hope to derive a generic quantitative statement concerning that point. Generally speaking, we expect that as L becomes larger and the subsample size smaller, the base method will receive less information and therefore will get closer to random guessing; in other words, we expect that for larger L , base selection probabilities $p_L(d)$ are all pulled together closer to the value q_L/D (corresponding random selection). Even assuming the ranking of the values of $p_L(d)$ is unchanged for different values

of L , the set $A_{\theta,L}^c$ of relevant variables at base selection probability larger than θ will contain a smaller number of covariates for larger L and fixed $\theta > q_L/D$. To maintain the same number of covariates in this relevant set, one has to consider a lower value of θ for larger L . In other words, as L increases, there is a trade-off between variance reduction of the selection frequencies $\Pi_L^{SFS}(d)$ as quantified by Theorem 1, and the reduced separation of their means $p_L(d)$, both of which are important for successful discrimination of relevant covariates. A similar effect occurs in L -fold cross-validation as when L grows the number of test sets available increases, but the size of each test set decreases. This trade-off has been discussed extensively, see for instance Arlot and Celisse (2010, Section 10.3).

Illustration of the bounds. In Figures 1 and 2 we illustrate the bound given by Corollary 1. On Figure 1, we show the dependence of the bound on τ for $D = 1000$, $q_L = 28$ and $L \in \{2, 4, 8\}$. We see that, for a fixed error level guaranteed by the bound, for larger L we can choose a smaller frequency selection threshold τ and therefore possibly select more covariates (while keeping in mind the tradeoff effect discussed in Remark 1). Moreover, we see that for error control ensuring a very low level of false positives, the bound for $L = 2$ is not applicable, while the bound for larger L is.

In Figure 2, we fix the expected number of false positives that we can tolerate, and use for each $L \in \{2, 4, 8\}$ and each $q_L \in \{1, \dots, 100\}$ Corollary 1 to determine the smallest selection frequency τ_{\min} that guarantees this bound. On this figure, we enforce at most one false positive on average, results for two and five FPs are given in the supplementary material. Reported is the behaviour

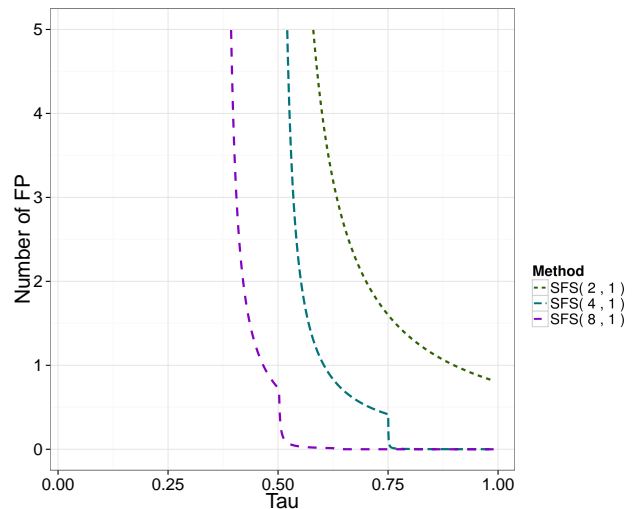


Fig. 1 Dependence of the bound of Corollary 1 on τ for various L , $q_L = 28$

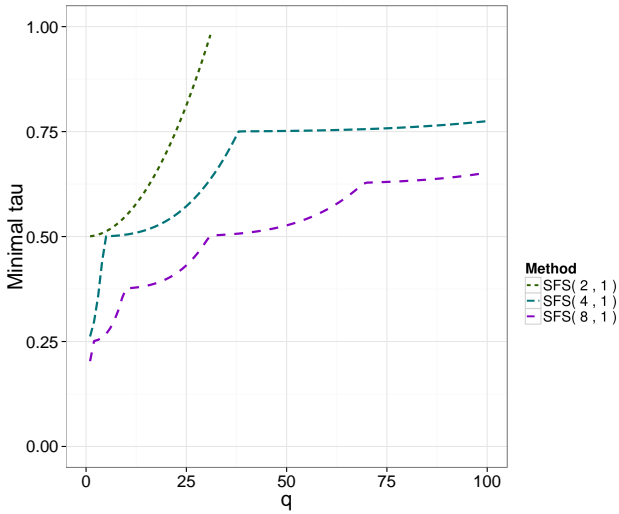


Fig. 2 The minimal value of τ such that the bound in Corollary 1 is below 1, as a function of q_L , for various L .

of τ_{\min} as a function q_L , the number of covariates selected by the baseline. Note that for standard stability selection ($L = 2$) the bound in Corollary 1 (which we recall coincides with the result of Meinshausen and Bühlmann 2010) does not achieve the value one for any $q_L \geq 32$.

An automatic choice of L is a thorny theoretical question and ultimately depends on the intended goal. If the goal is to select truly informative covariates with a strict control of the number of false positives, the above theoretical bounds can provide guidance; in Section 4.1, we investigate their practical relevance to drive the choice of L based on a given q_L and a target bound on the average number of false positives. If the goal is to improve prediction accuracy, we recommend indirect assessment methods such as cross-validation. If some constraints due to computing scalability are present, we recommend choosing L according to possible memory constraints or available parallel computing capabilities.

3.2 Randomization by taking covariate subsets

It is not possible to study the effect of taking random covariate subsets on the selection probabilities in as much generality as we studied the effect of observation subsampling in the previous section. The reason is that this effect depends prominently on the specific base procedure used (see Section 2.3). In this section, we assume an iterative, score-based base selection procedure (such as CMIM, see Section 2.3.1). We further limit the analysis and zoom in on a single iteration of this procedure (thus considering only the selection of 1 covariate) using a strongly simplified modelling of co-

variate scores. Although the result of further iterations generally strongly depends on previous ones, the prototypical model considered here for one iteration already highlights some interesting behaviour of the covariate subsampling procedure.

The simplified model is as follows. We assume that each covariate has an underlying score Q_d reflecting its true informativeness; only an estimation \hat{Q}_d of that score is available, which we assume to follow the simple additive model

$$\hat{Q}_d = Q_d + \varepsilon_d, \quad d = 1, \dots, D. \quad (10)$$

In Section 3.1 we used the probability $p_L(d)$ to be selected by the base method as a yardstick to tell signal from noise covariates. In that case, the quantity $p_L(d)$ is an example for a score Q_d .

We expect that the amplitude of the random estimation noise ε will typically depend on the size of the observations subsample and thus on the parameter L . However in this section we consider L as fixed and therefore omit it from the notation from now on.

The base procedure then outputs the covariate with the largest estimated score. We want to compare this base procedure to the *randomized base procedure* consisting in first picking at random a subset of $D' < D$ covariates, and returning out of those the one with the largest estimated score.

Similarly to what was considered in the previous section, we define uninformative covariates as those having true scores below a certain threshold θ :

$$A_{D,\theta} := \{1 \leq d \leq D : Q_d \leq \theta\}; \quad (11)$$

we also denote $A_{D,\theta}^c := \{1, \dots, D\} \setminus A_{D,\theta}$. Denote $p(d)$ and $p^{rand}(d)$ the probability of selecting covariate d using the deterministic and the randomized base procedure respectively. It is desirable for these to be as large as possible for $d \in A_{D,\theta}^c$, so that we compare the two base procedures by means of the sum of these probabilities, i.e., for the deterministic base procedure, $\sum_{d \in A_{D,\theta}^c} p(d) = \mathbb{P}[\hat{d}_D \in A_{D,\theta}^c]$, where \hat{d}_D denotes the index of the covariate selected by the deterministic base procedure.

In the following theorem we analyse the behaviour of the latter quantity as D grows. The main theoretical finding of this section is that, under certain circumstances concerning the distribution of the estimation noise, in an asymptotic sense \hat{d}_D will be determined only by the estimation error, and not by the true score. In other words, if D grows too large, the deterministic selection resembles picking at random.

This therefore supports the principle of randomizing the base procedure by taking subsets of covariates

in stability selection, since when the total number of covariates D is large enough, the probability of correct selection will be higher when taking a random covariate subset of size $D' < D$. This phenomenon is illustrated by a small simulation example at the end of the present section.

Theorem 2 *Consider a sequence of models of the form (10), a fixed number θ , and the following assumptions:*

- (i) *The true scores Q_d belong to the bounded interval $[0, M]$.*
- (ii) *The noise variables ε_i are independent and identically distributed, and their marginal distribution belongs, for some $\alpha > 0$, to the maximal domain of attraction (MDA) of a Fréchet(α) distribution (for details see Chapter 3.2 and 3.3 in Embrechts (1997)).*
- (iii) *As $D \rightarrow \infty$, $\frac{|A_{D,\theta}|}{D} \rightarrow \eta \in [0, 1]$, where $A_{D,\theta}$ is defined by (11).*

Then for $\hat{d}_D = \text{Arg Max}_{d=1,\dots,D} \hat{Q}_d$, it holds that

$$\lim_{D \rightarrow \infty} \mathbb{P} \left[\hat{d}_D \in A_{D,\theta} \right] \rightarrow \eta.$$

Interpretation of the theorem. Consider for comparison the blind strategy of simply drawing a covariate uniformly at random among D , regardless of observed scores. Then the probability for this covariate to lie in $A_{D,\theta}$ is obviously $|A_{D,\theta}|/D$. Thus, the theorem states that as D grows, the strategy consisting in picking the largest observed scores will asymptotically not be any better than the blind strategy.

Comments on assumption (iii). This assumption concerns the asymptotic repartition of the true scores Q_d , $1 \leq d \leq D$, as D grows. It is quite weak, and allows for the family of true scores to depend on D , provided this assumption remains satisfied. In particular, we can apply the theorem if the true scores are themselves random, and such that the assumption is satisfied a.s. In that situation, the theorem can be applied conditionally to the true score sequence and the conclusion will hold (for conditional incorrect selection probabilities) for almost all realizations of this sequence, and therefore also in expectation over the scores (i.e. for unconditional probabilities).

A simple and intuitive instance of the above is when the true scores Q_i are themselves assumed to be i.i.d. random draws following some (arbitrary) distribution on the interval $[0, M]$. In that case, assumption (iii) is satisfied a.s. with $\eta = \mathbb{P}[Q_1 \leq \theta]$ by the law of large numbers. Additionally, if the true scores are modelled as random i.i.d., the randomized procedure is equivalent to the base procedure with D simply replaced by D' . In this sense, the conclusion of the theorem applies to the randomized selection, as well. Denoting

$E_D := \mathbb{P}[\hat{d}_D \in A_{D,\theta}]$ in this setting, we have clearly $E_1 = \mathbb{P}[Q_1 \leq \theta] := \eta$ as well as $\lim_{D \rightarrow \infty} E_D = \eta$ by the theorem. On the other hand, it is easy to see that $E_D > \eta$ for any $D > 1$ (for any $D > 1$, selecting the covariate with largest observed score must be at least slightly better than random guessing). We conclude that E_D is not monotone in D , and that it must attain a minimum value for some finite $D_{\text{opt}} > 1$. (In the simulations shown at the end of the section, we actually see that E_D is unimodal.) The same applies to \hat{d}_D^{rand} , and we conclude that if $D > D_{\text{opt}}$, then it brings an advantage to select covariates out of random subsets of size smaller D' (the optimal size being $D' = D_{\text{opt}}$). To sum up the finding in an (apparently) paradoxical statement: as D grows too large, the deterministic selection behaves more randomly than the randomized selection using $D' < D$.

Comments on assumption (ii). The independence assumption is needed to apply classical results of extreme value theory. It is arguably unrealistic, and made here in order to illustrate the phenomenon in the simplest setting possible. We note that some extreme value results are also available under weak dependence models (Leadbetter et al., 1983, Chapter 3), so that this assumption might be relaxed somewhat, though this is out of the scope of the present work.

The assumption that the noise marginal distribution belongs to MDA(Fréchet) is needed to apply classical results of extreme value theory. Without getting into the details, this assumption (roughly) means that the distribution is heavy-tailed. This family includes Cauchy, Student's t, Pareto, Burr and Loggamma distributed noise (Embrechts, 1997, Table 3.4.2). This assumption is reasonable, for instance, if we consider that the estimated scores are based on t -statistics estimated from a limited number of observations. In contrast, the next result shows that the phenomenon described in Theorem 2 does not occur for Gaussian distributed noise:

Theorem 3 *Consider a sequence of models of the form (10), a fixed number θ , and the following assumptions:*

- (i) *the noise variables ε_i are independent and identically distributed with normal distribution.*
- (ii) *$\liminf_{D \rightarrow \infty} \frac{|A_{D,\theta}^c|}{D} > 0$, where $A_{D,\theta}$ is defined by Equation (11).*

Then for any $\theta' < \theta$:

$$\lim_{D \rightarrow \infty} \mathbb{P} \left[\hat{d}_D \in A_{D,\theta'} \right] = 0.$$

Remark. The above does not contradict existing results on support recovery of Lasso (reducing to a simple score thresholding procedure in the case of orthogonal

design), which is of sample complexity $O(\log(D))$ and thus eventually fails if the dimension D grows too large, for a fixed sample size. This is because such results study the exact recovery of all of the support covariates. By contrast, the result of Theorem 3 concerns the much weaker requirement of recovery (with probability close to 1) of one single informative covariate in a situation where the total proportion of informative covariates is non-vanishing.

We illustrate by a simulation study how the error probability $\mathbb{P}[\hat{d}_D \in A_{D,\theta}]$ depends on the distribution of the noise ε_d . We simulated 10000 noisy scores according to Equation (10) where $Q_d \sim 2 * \text{Bernoulli}(0.1)$ for each of the five noise distributions Gaussian, Cauchy, Student's t with 3, 5 and 10 degrees of freedom. We count how often the largest noisy score comes from an uninformative true score and plot this frequency against the total number of scores D in Figure 3. We see that for the Gaussian distribution, the probability to select a uninformative covariate is monotone decreasing to zero as D increases. In contrast, for heavy tailed distributions such as Cauchy and Student's t, the error probability decreases until reaching a minimum and increases afterwards.

We believe that the above results can be extended to the selection of k covariates $(\hat{d}_D^{(1)}, \dots, \hat{d}_D^{(k)})$ having the largest scores, using results about the joint limit distribution of k upper order statistics (Embrechts, 1997, Theorem 4.2.8). However, we emphasize that this does

not appear to be a direct consequence of the above result, since $\mathbb{P}[\hat{d}_D^{(k)} \in A_{D,\theta}]$ is not necessarily monotone increasing in k .

4 Experimental results

In this section we evaluate the performance of our method using three different criteria and several datasets. For the first two criteria, we aim at the identification of informative covariates and consider the number of true and false positives selected in a controlled framework where the ground truth is known. In the third setting, we consider a real-data classification task and evaluate the effect of variable selection on the prediction performance of a learning algorithm using only the selected covariates.

4.1 Identification of informative covariates

Setting. For the first two criteria we generate the output variable of interest according to a known linear model (except for the Vitamin dataset):

$$Y^{(n)} = \langle X^{(n)}, \beta \rangle + \varepsilon^{(n)} \quad \forall n = 1, \dots, N, \quad (12)$$

where $X^{(n)}$ and β are D -dimensional random vectors and $\varepsilon^{(n)} \sim \mathcal{N}(0, 1)$.

Additional experiments where $\varepsilon^{(n)}$ follows a heavy-tailed Student(3) distribution are given in the supplemental material only, as the results are similar. The vector of coefficients β contains only 20 non-zero entries. Their indices are chosen randomly and their values are generated from a $U[0, 1]$ distribution. This setting is similar to the one considered by Meinshausen and Bühlmann (2010).

We consider several different settings for the design matrix X , corresponding to controlled simulated situations or to real data. Except otherwise specified, each experiment is performed for $N = 500$ observations and $D = 1000$ covariates.

- 4 Blocks: The covariates are divided into 4 blocks with correlation inside but not among the blocks. The covariates follow a multivariate normal distribution $\mathcal{N}_D(0, \Sigma)$, where $\Sigma_{i,j} = 0.8 * \mathbf{1}\{i = j \bmod 4\}$
- Toeplitz design: The correlation between two covariates is higher the closer their indices are. The covariates follow a multivariate Normal distribution $\mathcal{N}_D(0, \Sigma)$, with $\Sigma_{i,j} = 0.99^{|i-j|}$
- Toeplitz (grouped predictors): As Toeplitz design, but the indices of the informative covariates consist of 5 groups of 4 indices, each drawn uniformly in the interval $[100g - 20, 100g + 20]$ where g is the

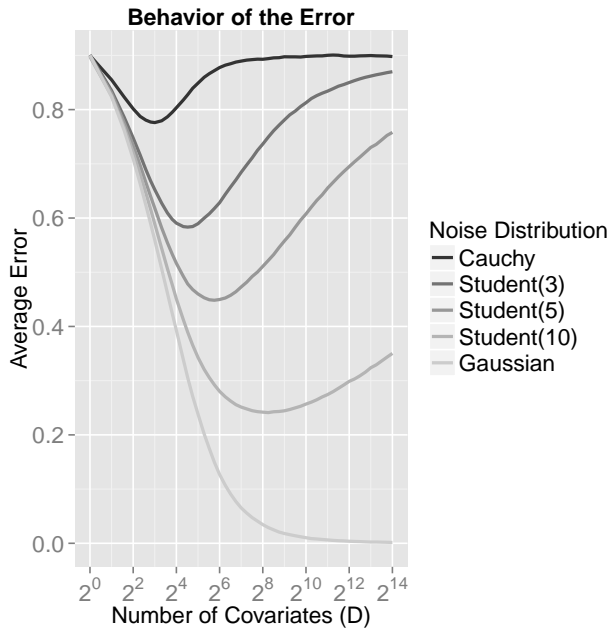


Fig. 3 Dependence of the average error probability on D for different noise distributions

group number. Therefore the informative covariates exhibit a cluster structure.

- 10 factors: Each covariate X_d is generated as a linear combination of unknown latent factors $X_d = \sum_{i=1}^{10} f_{d,i} \Phi_i + \nu_d \forall d = 1, \dots, D$ where the latent factors Φ_i , and the noise ν_d follow a standard normal distribution. The factor loading coefficients $f_{d,i}$ are fixed for any given realization of the dataset and are drawn beforehand from a $\mathcal{N}(0, 1)$ distribution.
- Correlated informative covariates, independent noise: The covariates follow a multivariate normal distribution $\mathcal{N}_D(0, \Sigma)$, with $\Sigma_{i,j} = 0.9$ for all indices i, j of informative covariates and $\Sigma_{i,j} = 0$ elsewhere.
- Vitamin dataset: A gene expression dataset considered by Meinshausen and Bühlmann (2010) (with $N = 115$, $D = 4088$.)

The only dataset that was not generated by the linear model is the Vitamin dataset. Following Meinshausen and Bühlmann (2010) we choose 6 covariates that have high correlation with the target as signal covariates. We permute all other covariates across the samples to break its dependence with the target. The covariance structure of these noise covariates is kept as we use the same permutation for all covariates.

For all datasets except the Vitamin dataset we adjust the signal to noise ratio $\text{Var}(\langle X^{(n)}, \beta \rangle) / \text{Var}(\varepsilon^{(n)}) = 2$ for the first criterion and 8 for the second and generate the vector Y by the linear model given in Equation (12).

Finally, we denote an instance of the proposed method by $SFS(L, V)$ with L, V as in Section 2.1.

4.1.1 Results with Precision@20 criterion

Experimental protocol. For the first criterion, we compare the number of truly informative covariates among the top $k = 20$ selected, when ranked by selection frequency. This criterion is also known as precision@k in the information retrieval literature. We compare the proposed methods for each design with several choices of parameters, the usual stability selection and the Lasso as reference methods¹. We consider the choices $SFS(L, 1)$ for $L \in \{2, 4, 8\}$ to investigate the effect of the subsample size of observations, and $SFS(2, V)$ for $V \in \{1, 2, 4, 8\}$ to investigate the effect of the amount of randomization of the base method by taking covariate subsets. Note that the $SFS(2, 1)$ is the standard stability selection.

We studied the performance of the different methods for a range of possible regularization parameters of

the base method. More specifically, rather than comparing the methods for a grid of fixed values of the regularization parameter λ of Lasso, see Equation (3), for each realization of the data, we used the values λ_{q_L} such that exactly $q_L \in \{1, \dots, 100\}$ covariates are selected by the base method. (In case several such values exist, the largest one is picked.) This approach seemed more fair and in line with the analysis of Section 3, in the sense that the number of covariates selected by the baseline, $q_L = \mathbb{E} [|S_{L,V}^{\text{base}}|]$, is kept constant across choices of (L, V) . For each value of q_L , we report the average precision@k, i.e., how many informative covariates are selected amongst the top $k = 20$ covariates ranked by their selection frequency. For standalone Lasso, we ranked the covariates by their estimated regression coefficients. The performance reported in Figure 4 and 5 is an average over 100 repetitions of each experiment.

Results. A first finding from Figure 4 is that standard stability selection ($SFS(2, 1)$) does not systematically outperform standalone Lasso, which appears to contradict the results of Meinshausen and Bühlmann (2010). A first reason for this is that here are some substantial differences between our evaluation criterion and theirs; this is discussed in more detail below. Perhaps more importantly, we evaluate the performance of standalone Lasso differently. Meinshausen and Bühlmann (2010) consider a single ranking of covariates by their selection order along the “Lasso path”, while we rank covariates by the magnitude of their estimated Lasso coefficients (for each value of q_L taken separately).

We find that using our ranking allows standalone Lasso to recover informative covariates more successfully when it first selects a total number of covariates higher than the true actual number of informative ones, then only keeps the ones with largest coefficient magnitude. In other words, relevant covariates might at first not be included in the selection set along the Lasso path, but once they are, they tend to get estimated coefficients larger than those of noisy covariates. As a consequence, in the regime where q_L is markedly larger than the true number of informative covariates, the success of stability selection over standalone Lasso is not systematic and appears to depend on the setting.

In additional experimental results presented in the supplemental material, we investigated the possibility of taking advantage of this effect by modifying the base method: We allowed Lasso to select 40 covariates, ranked them by the magnitude of their coefficients and kept a fixed number q_L (between 1 and 40) of covariates with largest coefficients. This is a version of the thresholded Lasso, that has been shown to have favourable asymptotic properties compared to the regular Lasso, see for instance Lounici (2008) and

¹ We used the R-package LARS (Hastie and Efron, 2012) as Lasso implementation.

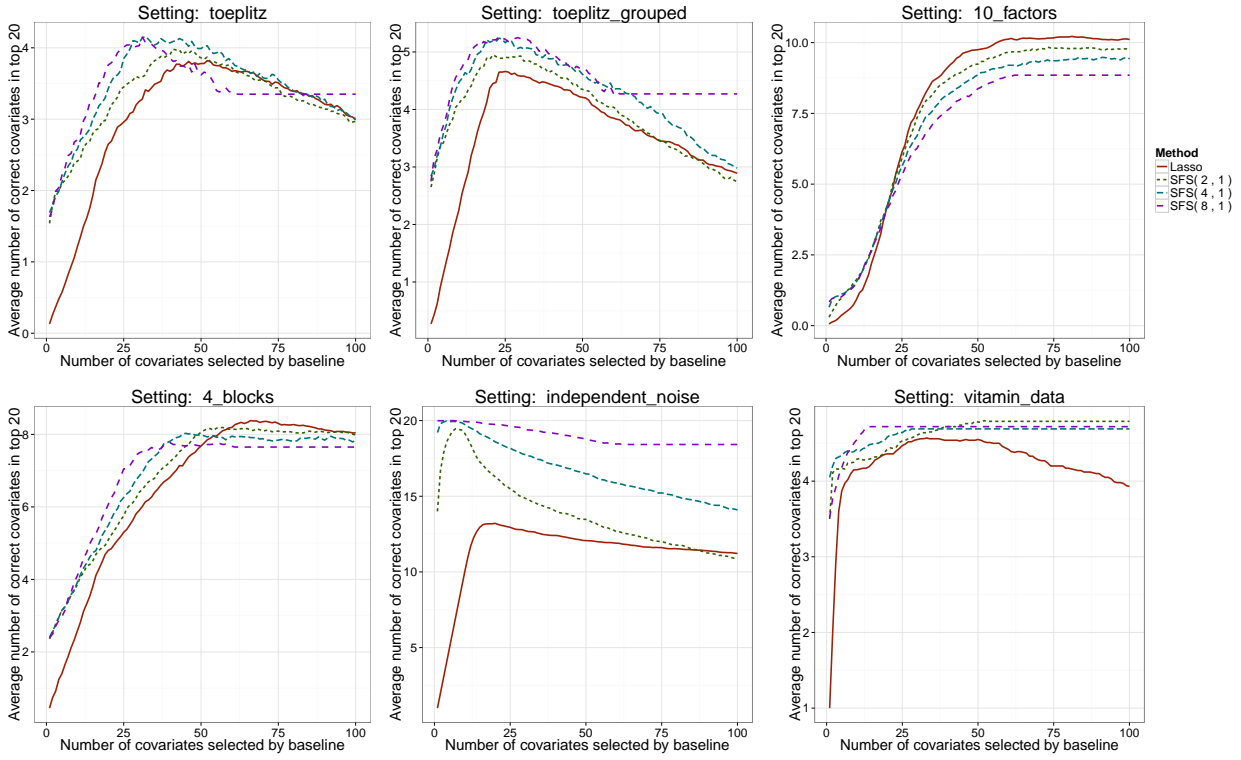


Fig. 4 Comparison of plain Lasso and stability selection with varying numbers of disjoint observation subsamples (corresponding to line types). We plot the average number of informative covariates among the top 20 scored (for the Vitamin dataset: top 6), depending on the number of covariates selected by the base method, figures correspond to designs. Each dataset contains 20 signal and 980 noise covariates, except for the Vitamin dataset, which contains 6 signal and 4082 noise covariates.

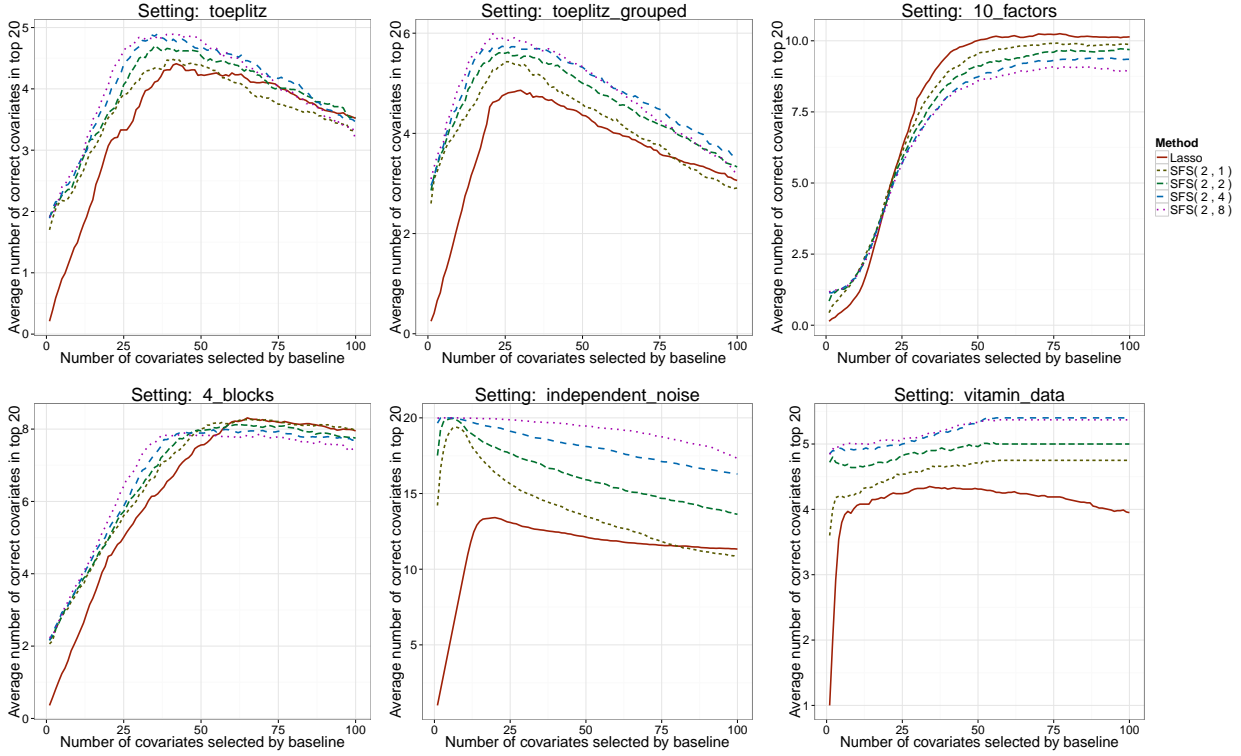


Fig. 5 Like Figure 4, except that we fix the number of disjoint observation subsamples and vary the number of disjoint covariate subsets.

Meinshausen and Yu (2009). However, we found that the precision@20 results of applying SFS on top of this base method are remarkably stable and unaffected by the value of q_L in this setting. Thus, thresholded Lasso and SFS does not improve significantly over regular Lasso and SFS.

Second, we observed in Figure 4 and 5 that in most cases if for a fixed value of q_L standard stability selection $SFS(2,1)$ improves on plain Lasso, then the extended methods we propose generally lead to a further improvement. A qualitative observation is that stability selection appears to be more successful for situations with limited or short-range dependence (Toeplitz design, Toeplitz grouped, independent noise, realistic datasets with presumably weak dependence) than with systematic dependence structure (factorial design, block design). Naturally, and as expected from the analysis and discussion in Section 3, eventually the performance deteriorates again if the size of the observation subsamples N/L becomes too small to allow a reasonable estimation, or if the number of covariates D/V used in the randomized base procedure is too low (experiments not shown). Overall, these results suggest to use extended stability selection, both for reasons of potential increase in performance, and of efficient scalability via possible parallelization over standard stability selection, whenever the latter itself improves on the base method.

Discussion of the precision@20 criterion. In their original work Meinshausen and Bühlmann (2010) reported two evaluation measures: the first was the probability that the top ℓ covariates by selection frequency are all relevant (for some fixed beforehand ℓ). The second was the number of noise variables included in the top- ℓ selection, where ℓ is chosen (separately in each data realization) so that the selection contains at least a fixed proportion such as 20% of the true variables. In experiments using these criteria, we found that the first criterion was only poorly informative, in the sense that it was often very close to either 0 or 1, not providing a very clear contrast between the methods; and that the second criterion was in many cases subject to a very large expectation and variance, so that we were also wary about its relevance. For this reason, we chose precision@20 as the performance measure; this criterion has the advantage of being stable, comparable across settings, in plausible relation to intended applications, and is standard in information retrieval. Additionally, for stability selection Meinshausen and Bühlmann (2010) only reported results for a specific value of q_L , namely $q_L \approx \sqrt{0.8D}$, and we preferred to report performance over a range of values of q_L . Overall we believe our protocol presents a more complete and fairer overall picture.

4.1.2 Results with theoretical control of false positives

In this section we first illustrate and validate the bound given by Corollary 1.

We now perform experiments similar to those described above, to investigate the effect of observation subsampling while using the value of τ dictated by Corollary 1 in order to achieve less than one expected false positive (i.e. the values represented in Figure 2). We determine the average number of false positives for $q_L \in \{1, \dots, 100\}$. After comparing the empirical number of false positives to the theoretical bound, we also count the number of true positives to determine the power of the method. As we noticed that for a signal to noise ratio of 2 hardly any covariates are selected at all, we used a signal to noise ratio of 8. (In absence of a false positive error bound for standalone Lasso, the latter is not included in these experiments.) The results are given in Figures 6 and 7. The main findings concerning the comparison of standard stability selection and extended stability selection are the following:

- The bound for the number of false positives holds in all experiments except for factorial and grouped Toeplitz design when q_L is small; it seems to indicate that the symmetry assumptions of Corollary 1 are significantly violated in these situations.
- Extended stability selection selects fewer false positives in all experiments.
- Extended stability selection selects more true positives in all experiments except for the Vitamin dataset.
- The largest number of true positives is often achieved for some q_L that is larger than the number of relevant covariates and often lies outside of the regime where the bound for standard stability selection can be used.

We observe in Figure 6 that the FP bound for extended stability selection seems to be loose. There is probably room for improvement in the theoretical bounds, for instance using ideas from Shah and Samworth (2013) to improve on Markov's inequality under additional assumptions on the distribution of the frequency counts, though this is outside of the intended scope of this paper.

4.2 Application: Image classification.

In this section, we use variable selection as a preprocessing step to solve a classification problem. We investigate whether stabilising the variable selection method with extended stability selection improves the classification error rate compared with applying the variable selection method to the whole dataset.

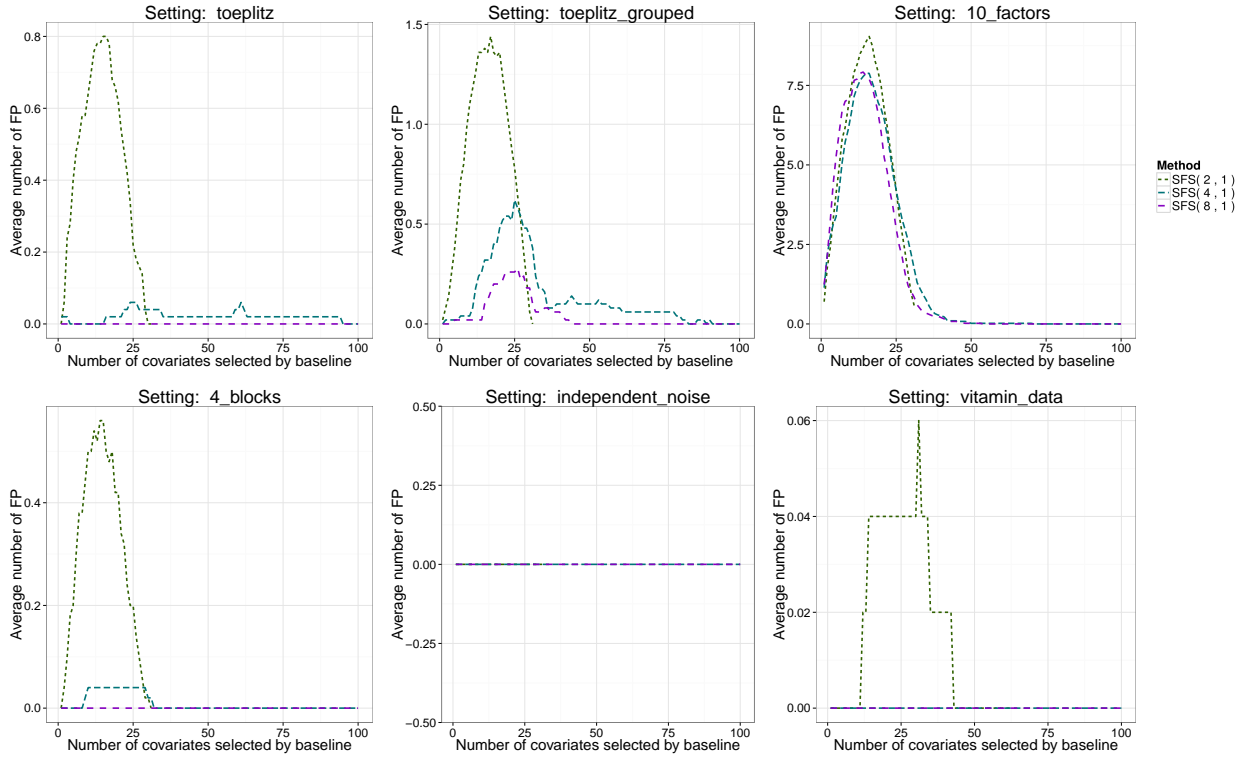


Fig. 6 Comparison of stability selection and its extensions for various numbers of observation subsamples (corresponding to line types), we choose the largest τ such that in Corollary 1 the expected number of false positives is less than one. We plot the average number of false positives, depending on the number of covariates selected by the base method, figures correspond to designs. Each dataset contains 20 signal and 980 noise covariates, except for the Vitamin dataset, which contains 6 signal and 4082 noise covariates.

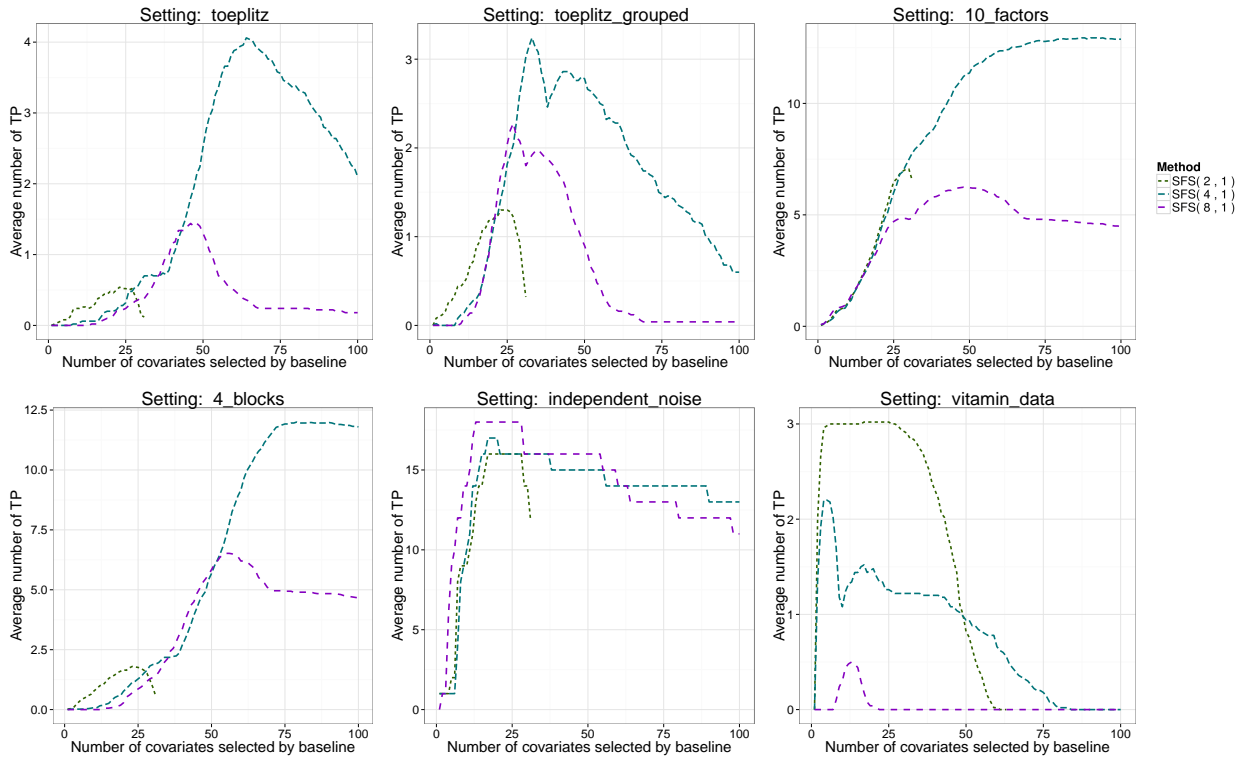


Fig. 7 Like Figure 6, except that we plot the average number of true positives.

Dataset Description. For prediction we used a subset of 6000 images of handwritten digits from the MNIST dataset (LeCun et al., 1998). Covariates were computed using feature extractors from the a collaborative machine learning framework (MASH Consortium, 2012) where external contributors can directly submit feature extractors of their choice by uploading C++ code through a web interface. A total of 48416 covariates were computed on each image from the contributed feature extractors. These covariates are heterogeneous because they come from different contributors, generally exhibit strong correlations, do not have a sparse structure, and many of them contain at least some information about the class label. Here, the goal is not to strictly identify informative covariates, but to reduce the dimensionality of the problem in order to apply a learning and prediction method afterwards. This way, the computational complexity and memory requirements of the learning algorithm can be reduced considerably. Note that in these experiments we are not aiming at outperforming state-of-the-art classification results on MNIST, but we wish to evaluate the effect of the proposed extensions to stability selection on prediction performance, as compared to standard stability selection with the same base method.

Experimental setup and results. As base method, we use CMIM (conditional mutual information maximization, Fleuret, 2004) to select 1000 covariates. As described in Section 2.3, Equations (1) and (2) CMIM iteratively selects covariates updating in each iteration a score for each covariate. To speed up computation, we use an approximation proposed by Beinrucker et al. (2012a) and only update scores up until iteration k (re-computation of the scores at each iteration being the computationally costly part), so that Equation (2) is replaced by

$$\nu(\ell) = \underset{d \in \{1, \dots, D\} \setminus \{\nu(i), i < \ell\}}{\text{Arg Max}} \min_{j \leq \min(\ell-1, k)} \hat{I}(X_d; Y | X_{\nu(j)}).$$

In other words, after k normal CMIM iterations the $1000 - k$ covariates which remain to be selected are chosen as the ones with the largest values of the score $\min_{j \leq k} \hat{I}(X_d; Y | X_{\nu(j)})$. We considered $k = 10$ and $k = 100$ and denoted the resulting stability selection method as SFS_{10} or SFS_{100} respectively. Selection by CMIM only (updating scores until iteration 10) is denoted by CMIM_{10} .

For all considered methods 1000 covariates are selected at the end, corresponding to about 2.1% of the total number of covariates. Using the selected covariates we applied the AdaBoost.MH algorithm (Escudero et al., 2000; Schapire and Singer, 1999) as a learning and

prediction algorithm with various numbers of boosting iterations. We report our results in Table 1 and 2. In Table 2 we only show results for SFS_{10} , results for SFS_{100} were similar. The main conclusions are the following:

- We observe a slight trend that $L > 2$ leads to better prediction performance than $L = 2$, but it is not statistically significant. The main conclusion here though, is that taking smaller subsample sizes does not degrade the final performance. This is of much relevance in practice, since smaller subsample sizes require less memory. Furthermore, if the computational complexity $C_{\text{Base}}(N)$ of the base method grows faster than linearly in the number of observations N , the total computing cost is also reduced (since $LC_{\text{Base}}(N/L)$ is decreasing in L in that case). Parallelization is also easier for smaller disjoint subsamples.
- We observe a statistically significant trend that $V > 1$ (smaller covariate subsets) leads to improvements

Table 1 The effect of the size of covariate subsets on prediction error. Selection of 1000 covariates with $\text{SFS}_k(2, V)$ using CMIM_k as base method; prediction with AdaBoost.MH with various numbers of iterations. Reported are averaged results over 100 repetitions.

# it.	$\text{SFS}_{10}(2, V), V =$			
	1	2	6	10
50	8.1 (0.2)	7.7 (0.3)	7.5 (0.3)	7.4 (0.3)
100	5.0 (0.2)	4.8 (0.2)	4.5 (0.1)	4.5 (0.3)
200	3.3 (0.1)	2.9 (0.1)	2.5 (0.2)	2.7 (0.2)
400	2.8 (0.1)	2.7 (0.2)	2.4 (0.1)	2.4 (0.2)
800	2.4 (0.1)	2.5 (0.2)	2.2 (0.1)	2.1 (0.1)
1600	2.0 (0.1)	2.0 (0.2)	1.9 (0.2)	1.8 (0.1)
# it.	$\text{SFS}_{100}(2, V), V =$			
	1	2	6	10
50	8.0 (0.2)	7.5 (0.3)	7.3 (0.2)	7.2 (0.3)
100	4.9 (0.2)	4.6 (0.2)	4.0 (0.3)	3.0 (0.2)
200	3.3 (0.2)	3.0 (0.2)	2.3 (0.1)	2.6 (0.2)
400	3.0 (0.1)	2.5 (0.1)	2.1 (0.1)	2.2 (0.1)
800	2.3 (0.1)	2.2 (0.1)	2.0 (0.1)	1.9 (0.1)
1600	2.1 (0.2)	1.9 (0.1)	1.8 (0.1)	1.6 (0.1)

Table 2 The effect of the size of observation subsamples on prediction error. Selection of 1000 covariates with CMIM_{10} only and $\text{SFS}_{10}(L, 1)$ using CMIM_{10} as base method; prediction with AdaBoost.MH with various numbers of iterations. Reported are averaged results over 100 repetitions.

# it.	CMIM_{10}	$\text{SFS}_{10}(L, 1), L =$		
		2	6	10
50	8.4 (0.4)	8.1 (0.2)	8.0 (0.3)	8.1 (0.2)
100	5.5 (0.2)	5.0 (0.2)	5.2 (0.3)	4.9 (0.2)
200	3.6 (0.3)	3.3 (0.1)	3.2 (0.2)	3.4 (0.2)
400	2.8 (0.2)	2.8 (0.1)	2.7 (0.2)	2.7 (0.1)
800	2.4 (0.1)	2.4 (0.1)	2.3 (0.2)	2.3 (0.1)
1600	2.1 (0.1)	2.0 (0.1)	2.1 (0.1)	1.9 (0.1)

in prediction performance in comparison to $V = 1$ (standard stability selection).

Overall, and paralleling the conclusions of Section 4.1, these results demonstrate the relevance of extended stability selection.

5 Discussion/Conclusion

We presented theoretical and experimental support for the proposed extensions of the stability selection methodology. Concerning subsampling of observations using smaller subsamples, we generalized error bounds of preceding investigations (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013). These new bounds give insights into the effect of the subsample size on the selection performance and can provide guidance to practitioners to use a version of stability selection that suits their needs. Our simulations showed that the obtained false positive bounds can apply in regimes where previous bounds are void, resulting in a more powerful procedure (number of true positives). Still, the obtained bounds appear loose in practice and could probably be further refined. Concerning randomising of an arbitrary base procedure by taking random covariate subsets, we motivated this method from a simplified theoretical toy model, showing in certain circumstances (heavy tailed score noise) that restricting the search to a random subset increases the probability of correct recovery. We expect that this second type of extension should be particularly appealing for practitioners in high dimensional settings, where the number of covariates largely exceeds the number of observations.

Experimental results using the precision@20 information retrieval criterion showed that both extensions improve on standard stability selection whenever the latter improves on the base method applied to the whole dataset. It remains an open task to determine precise conditions under which a variable selection method can be improved by stability selection. Even though our analysis gave first insights on the dependence of the error probability on the size of observation subsamples, a more precise rule for the optimal choice of the subsample size is left for further work. The same holds for the optimal choice of the size of covariate subsets, even though in practice in very high dimensional problems this choice might be dictated by computational constraints in the first place.

Acknowledgements We are extremely grateful to Nicolai Meinshausen and Peter Bühlmann for communicating to us the R-code used by Meinshausen and Bühlmann (2010) as well as for numerous discussions. We are indebted to Richard Samworth and Rajen Shah for numerous discussions and for

hosting the first author during part of this work. We thank Maurilio Gutzeit for helping us with part of the numerical experiments.

A Proofs of theoretical results

A.1 Proofs of Section 3.1

For notational convenience we use the shorthand $S^{\text{base}}(\ell, t) \equiv S^{\text{base}}(X^{(S(\ell, t))}, Y^{(S(\ell, t))})$. To prove Theorem 1 and Corollary 1 we need some notation and two lemmas. We define

$$\Pi_{L, \ell_0}^{\text{simult}}(d) := \frac{1}{T} \sum_{t=1}^T \mathbf{1} \left\{ \sum_{\ell=1}^L \mathbf{1} \{d \in S_L^{\text{base}}(\ell, t)\} \geq \ell_0 \right\}$$

the ratio of repetitions out of T where covariate d has been selected in at least ℓ_0 subsamples simultaneously.

Lemma 1 (*Relation of Π^{simult} and Π^{FS}*) *It holds for any $d \in \mathcal{F}$:*

$$\left(\frac{L - \ell_0 + 1}{L} \right) \Pi_{L, \ell_0}^{\text{simult}}(d) + \frac{\ell_0 - 1}{L} \geq \Pi_L^{\text{FS}}(d).$$

Proof. We have for all repetitions of drawings of subsamples $t = 1, \dots, T$:

$$\begin{aligned} & \frac{1}{L} \sum_{\ell=1}^L \mathbf{1} \{d \in S^{\text{base}}(\ell, t)\} \\ & \leq \left(\frac{\ell_0 - 1}{L} \right) \mathbf{1} \left\{ \sum_{\ell=1}^L \mathbf{1} \{d \in S^{\text{base}}(\ell, t)\} \leq \ell_0 - 1 \right\} \\ & \quad + \mathbf{1} \left\{ \sum_{\ell=1}^L \mathbf{1} \{d \in S^{\text{base}}(\ell, t)\} \geq \ell_0 \right\}. \end{aligned}$$

Averaging over the repetitions $t = 1, \dots, T$, we obtain

$$\begin{aligned} \Pi_L^{\text{FS}}(d) & \leq \frac{\ell_0 - 1}{L} (1 - \Pi_{L, \ell_0}^{\text{simult}}(d)) + \Pi_{L, \ell_0}^{\text{simult}}(d) \\ & = \left(\frac{L - \ell_0 + 1}{L} \right) \Pi_{L, \ell_0}^{\text{simult}}(d) + \frac{\ell_0 - 1}{L}. \end{aligned}$$

□

Lemma 2 (*Exponential inequality for Π^{simult}*) *The following inequality holds for any $d \in \mathcal{F}$, $\xi > 0$, and $\ell_0 \in \{1, \dots, L\}$ such that $p_0 := \frac{\ell_0}{L} \geq p_L(d)$:*

$$\mathbb{P} [\Pi_{L, \ell_0}^{\text{simult}}(d) \geq \xi] \leq \frac{1}{\xi} \exp(-LD(p_0, p_L(d))). \quad (13)$$

Proof. We have

$$\begin{aligned} \mathbb{E} [\Pi_{L, \ell_0}^{\text{simult}}(d)] & = \mathbb{P} \left[\sum_{\ell=1}^L \mathbf{1} \{d \in S^{\text{base}}(\ell, 1)\} \geq \ell_0 \right] \\ & = \mathbb{P} [\text{Bin}(L, p_L(d)) \geq \ell_0] \\ & \leq \exp(-LD(p_0, p_L(d))). \end{aligned}$$

The first equality is valid because the L random observation subsamples are *disjoint*. Therefore, their joint distribution is the same as that of L independent samples of size $\lfloor \frac{N}{L} \rfloor$; thus $(S^{\text{base}}(\ell, 1))_{1 \leq \ell \leq L}$ has the same distribution as L independent copies of the variable S_L^{base} . The last inequality is the Chernoff binomial bound. Using Markov's inequality we get (13). □

Proof of Theorem 1. We relate Π^{SFS} to Π^{simult} and apply an exponential inequality on Π^{simult} . For any $d \in A_{\theta,L}$, it holds by definition of $A_{\theta,L}$ and the assumptions on p_0 that $p_L(d) \leq \theta \leq p_0$, hence it holds by Lemma 1 and Lemma 2 that

$$\begin{aligned} \mathbb{P}[\Pi_L^{SFS}(d) \geq \tau] &\leq \mathbb{P}\left[\left(\frac{L - \ell_0 + 1}{L}\right) \Pi_{L,\ell_0}^{\text{simult}}(d) + \frac{\ell_0 - 1}{L} \geq \tau\right] \\ &= \mathbb{P}\left[\Pi_{L,\ell_0}^{\text{simult}}(d) \geq \frac{L\tau - \ell_0 + 1}{L - \ell_0 + 1}\right] \\ &\leq \frac{1 - p_0 + L^{-1}}{\tau - p_0 + L^{-1}} \exp(-LD(p_0, p_L(d))), \end{aligned}$$

where we have used $\xi := \frac{L\tau - \ell_0 + 1}{L - \ell_0 + 1}$. This result generalizes Shah and Samworth (2013, Lemma 5). Hence

$$\begin{aligned} \frac{\mathbb{E}[|S_{L,\tau}^{SFS} \cap A_{\theta,L}|]}{|A_{\theta,L}|} &= \frac{1}{|A_{\theta,L}|} \sum_{d \in A_{\theta,L}} \mathbb{P}[\Pi_L^{SFS}(d) \geq \tau] \\ &\leq \frac{1 - p_0 + L^{-1}}{\tau - p_0 + L^{-1}} \frac{1}{|A_{\theta,L}|} \sum_{d \in A_{\theta,L}} \exp(-LD(p_0, p_L(d))). \end{aligned}$$

Since $x \rightarrow \exp(-LD(p_0, x))$ is non-decreasing, we obtain the first part of the result by upper bounding for all $d \in A_{\theta,L}$:

$$\exp(-LD(p_0, p_L(d))) \leq \exp(-LD(p_0, \theta)).$$

For the second part, we use the upper bound

$$\begin{aligned} \exp(-LD(p_0, p_L(d))) &= \frac{\exp(-LD(p_0, p_L(d)))}{p_L(d)} p_L(d) \\ &\leq \frac{\exp(-LD(p_0, \theta))}{\theta} p_L(d), \end{aligned}$$

since the function $x \mapsto \frac{\exp(-LD(p_0, x))}{x}$ can be shown to be non-decreasing for $x \leq p_0 - L^{-1}$. Finally, summing over $d \in A_{\theta,L}$, observe

$$\sum_{d \in A_{\theta,L}} p_L(d) = \mathbb{E}\left[\sum_{d \in A_{\theta,L}} \mathbf{1}\{d \in S_L^{\text{base}}\}\right] = \mathbb{E}[|A_{\theta,L} \cap S_L^{\text{base}}|],$$

leading to the desired conclusion. Equations (6) and (7) can be proved similarly. \square

Proof of Corollary 1. This follows the same argument as in Shah and Samworth (2013). If the variable selection was completely at random, the marginal selection probability of any given covariate would be $\frac{q_L}{D}$, where we recall $q_L = \mathbb{E}[|S_L^{\text{base}}|]$ is the average number of covariates selected by the base method. As we assume that the selection probability of a signal covariate is better than random; it entails that for any $d \in \mathcal{N}^C$, we must have $p_L(d) > \frac{q_L}{D}$. Conversely, as all noise covariates have the same probability to be selected by the base method, one has $p_L(d) < \frac{q_L}{D}$ for any $d \in \mathcal{N}$. Therefore, with $\theta := \frac{q_L}{D}$ we must have $A_{\theta,L} = \mathcal{N}$ and $A_{\theta,L}^c = \mathcal{N}^C$. Inequality (4) therefore implies (8), wherein we have taken a minimum over the range of ℓ_0 allowed in Theorem 1. \square

A.2 Proofs for Section 3.2

Proof of Theorem 2. We can first bound the error probability from above by omitting Q_d :

$$\begin{aligned} \mathbb{P}[\hat{d}_D \in A_{D,\theta}] &= \mathbb{P}\left[\max_{d \in A_{D,\theta}} \hat{Q}_d > \max_{d \in A_{D,\theta}^c} \hat{Q}_d\right] \\ &= \mathbb{P}\left[\max_{d \in A_{D,\theta}} (Q_d + \varepsilon_d) > \max_{d \in A_{D,\theta}^c} (Q_d + \varepsilon_d)\right] \\ &\leq \mathbb{P}\left[\max_{d \in A_{D,\theta}} (\theta + \varepsilon_d) > \max_{d \in A_{D,\theta}^c} (\theta + \varepsilon_d)\right] \\ &= \mathbb{P}\left[\text{Arg Max}_{d \in \{1, \dots, D\}} \varepsilon_d \in A_{D,\theta}\right] = \frac{|A_{D,\theta}|}{D} \rightarrow \eta, \end{aligned}$$

as $D \rightarrow \infty$. If $\eta = 0$, the conclusion is therefore established; in the remainder of the proof we hence assume $\eta > 0$. We defer to the end of the proof the case $\eta = 1$ and assume for now that $\eta \in (0, 1)$. Then $\frac{|A_{D,\theta}|}{D} \rightarrow \eta \in (0, 1)$ implies both $|A_{D,\theta}| \rightarrow \infty$ and $|A_{D,\theta}^c| \rightarrow \infty$, as well as $\frac{|A_{D,\theta}^c|}{|A_{D,\theta}|} \rightarrow \gamma := \frac{1-\eta}{\eta}$. We return to the error probability and bound it from below by using $Q_d \geq 0$ for $d \in A_{D,\theta}$ and $Q_d \leq M$ for $d \in A_{D,\theta}^c$:

$$\mathbb{P}[\hat{d}_D \in A_{D,\theta}] \geq \mathbb{P}\left[\max_{d \in A_{D,\theta}} \varepsilon_d > M + \max_{d \in A_{D,\theta}^c} \varepsilon_d\right]. \quad (14)$$

Since the distribution of ε_i belongs to MDA($\text{Fréchet}(\alpha)$), from classical results of extreme value theory (Embrechts, 1997, Theorem 3.3.7) we know that there exists a slow varying function L so that, if we denote $G(x) := x^{1/\alpha} L(x)$, then

$$\begin{aligned} \frac{\max_{d \in A_{D,\theta}} \varepsilon_d}{G(|A_{D,\theta}|)} &\rightarrow \text{Fréchet}(\alpha) \\ \text{and} \\ \frac{\max_{d \in A_{D,\theta}^c} \varepsilon_d}{G(|A_{D,\theta}^c|)} &\rightarrow \text{Fréchet}(\alpha), \end{aligned} \quad (15)$$

in the sense of convergence in distribution, as $D \rightarrow \infty$. Following on (14):

$$\begin{aligned} \mathbb{P}[\hat{d}_D \in A_{D,\theta}] &\geq \mathbb{P}\left[\frac{\max_{d \in A_{D,\theta}} \varepsilon_d}{G(|A_{D,\theta}|)} > \frac{M}{G(|A_{D,\theta}|)} \right. \\ &\quad \left. + \frac{G(|A_{D,\theta}^c|)}{G(|A_{D,\theta}|)} \frac{\max_{d \in A_{D,\theta}^c} \varepsilon_d}{G(|A_{D,\theta}^c|)}\right]. \end{aligned}$$

As L is slowly varying, we have $\lim_{x \rightarrow \infty} \frac{L(ax)}{L(x)} \rightarrow 1$ uniformly for a belonging to a bounded interval of the positive real axis (Embrechts, 1997, Theorem A 3.2). We deduce

$$\frac{G(|A_{D,\theta}^c|)}{G(|A_{D,\theta}|)} = \left(\frac{|A_{D,\theta}^c|}{|A_{D,\theta}|}\right)^{\frac{1}{\alpha}} \frac{L\left(\frac{|A_{D,\theta}^c|}{|A_{D,\theta}|}\right)}{L(|A_{D,\theta}|)} \rightarrow \gamma^{1/\alpha}, \quad (16)$$

as $D \rightarrow \infty$. We apply Slutsky's theorem (Embrechts, 1997, Example A 2.7) to Equations (16) and (15) to obtain

$$\frac{G(|A_{D,\theta}^c|)}{G(|A_{D,\theta}|)} \frac{\max_{d \in A_{D,\theta}^c} \varepsilon_d}{G(|A_{D,\theta}^c|)} \rightarrow \text{Fréchet}(\alpha, \gamma^{1/\alpha})$$

in distribution, where $\text{Fréchet}(\alpha, c)$ denotes the $\text{Fréchet}(\alpha)$ distribution rescaled by a factor $c > 0$.

Further, slow variation of L implies that $L(x)$ is asymptotically negligible with respect to any power function, so that

$$\frac{M}{G(|A_{D,\theta}|)} = \frac{M}{|A_{D,\theta}|^{\frac{1}{\alpha}} L(|A_{D,\theta}|)} \rightarrow 0, \quad \text{as } D \rightarrow \infty. \quad (17)$$

As the maxima in $\max_{d \in A_{D,\theta}} \varepsilon_d$ and $\max_{d \in A_{D,\theta}^c} \varepsilon_d$ are taken over disjoint sets of independent random variables, they are independent. Since they converge marginally in distribution, they also converge jointly and their difference converges due to the continuous mapping theorem (Embrechts, 1997, Theorem A.2.6). Combining with (17) and using Slutsky's theorem again, we conclude that

$$\frac{\max_{d \in A_{D,\theta}} \varepsilon_d}{G(|A_{D,\theta}|)} - \frac{M}{G(|A_{D,\theta}|)} - \frac{G(|A_{D,\theta}^c|)}{G(|A_{D,\theta}|)} \frac{\max_{d \in A_{D,\theta}^c} \varepsilon_d}{G(|A_{D,\theta}^c|)}$$

converges in distribution to the difference of two independent Fréchet distributed random variables. This convergence implies the convergence of the c.d.f. for all continuity points (Embrechts, 1997, Equation A.1). As the limiting distribution is continuous, we finally obtain

$$\liminf_{D \rightarrow \infty} \mathbb{P}[\hat{d}_D \in A_{D,\theta}] \geq \mathbb{P}[F - \gamma^{\frac{1}{\alpha}} F' > 0],$$

where F, F' are independent Fréchet(α) random variables. To identify the value of this lower bound, observe that it is also the limiting value of

$$\mathbb{P}\left[\operatorname{Arg Max}_{d \in \{1, \dots, D\}} \varepsilon_d \in A_{D,\theta}\right] = \frac{|A_{D,\theta}|}{D}.$$

Indeed, it suffices to repeat the above argument, except for skipping inequality (14). Hence this limiting value is exactly equal to η .

Finally, for the case $\eta = 1$, observe that the above argument remains valid provided $|A_{D,\theta}^c| \rightarrow \infty$. Even if this is not the case (i.e. $|A_{D,\theta}^c|$ remains bounded), then the conclusion would be *a fortiori* true since we could replace $A_{D,\theta}^c$ by a slightly larger set of cardinality $\ln(D)$ (say), which can only decrease the lower bound while still obtaining the above limiting value. \square

Proof of Theorem 3. To show the convergence of the error probability we use similar arguments as in the proof of Theorem 2. From classical results of extreme value theory for independent standard normal random variables $(\zeta_k)_{k \in \mathbb{N}}$ (Embrechts, 1997, Example 3.3.29) it holds that

$$\frac{\max_{i \leq k} \zeta_i - b_k}{a_k} \rightarrow \text{Gumbel}, \quad \text{as } k \rightarrow \infty,$$

in distribution, where $a_k := \frac{1}{\sqrt{2 \ln(k)}}$ and $b_k := \sqrt{2 \ln(k)} - \frac{\ln(4\pi) + \ln(\ln(k))}{2\sqrt{2 \ln(k)}}$. Below, to clarify the argument we will introduce $(\zeta_k)_{k \in \mathbb{N}}$ and $(\zeta'_k)_{k \in \mathbb{N}}$ two independent sequences of independent standard normal variables.

Denote $k_D := |A_{D,\theta'}|$, $\ell_D := |A_{D,\theta}^c|$ and $\Delta := \theta - \theta' > 0$. Since $A_{D,\theta}^c \subseteq A_{D,\theta'}^c$, we can bound the error probability from above as follows:

$$\begin{aligned} \mathbb{P}[\hat{d}_D \in A_{D,\theta'}] &\leq \mathbb{P}\left[\max_{d \in A_{D,\theta'}} (Q_d + \varepsilon_d) > \max_{d \in A_{D,\theta}^c} (Q_d + \varepsilon_d)\right] \\ &\leq \mathbb{P}\left[\max_{d \in A_{D,\theta'}} \varepsilon_d > \max_{d \in A_{D,\theta}^c} \varepsilon_d + \Delta\right] \\ &= \mathbb{P}\left[\max_{d \leq k_D} \zeta'_d > \max_{d \leq \ell_D} \zeta_d + \Delta\right]. \end{aligned}$$

The last equality holds since $A_{D,\theta'}$ and $A_{D,\theta}^c$ are disjoint; it is purely formal but notationally convenient for the sequel. Now denote $k'_D := \max(k_D, \ell_D)$. The above implies

$$\begin{aligned} \mathbb{P}[\hat{d}_D \in A_{D,\theta'}] &\leq \mathbb{P}\left[\max_{d \leq k'_D} \zeta'_d > \max_{d \leq \ell_D} \zeta_d + \Delta\right] \\ &= \mathbb{P}\left[\frac{a_{k'_D}}{a_{\ell_D}} \left(\max_{d \leq k'_D} \zeta'_d - b_{k'_D}\right) - \frac{\max_{d \leq \ell_D} \zeta_d - b_{\ell_D}}{a_{\ell_D}} \right. \\ &\quad \left. > \frac{b_{\ell_D} - b_{k'_D} + \Delta}{a_{\ell_D}}\right] \end{aligned}$$

We treat the different terms in the above upper bound. We have

$$\frac{a_{k'_D}}{a_{\ell_D}} = \frac{\sqrt{2 \ln(\ell_D)}}{\sqrt{2 \ln(k'_D)}} \leq 1.$$

Noting that $b_k = \sqrt{2 \ln k} + o(1)$, we have

$$\begin{aligned} \frac{b_{\ell_D} - b_{k'_D} + \Delta}{a_{\ell_D}} &= \sqrt{2 \ln \ell_D} \left(\sqrt{2 \ln \ell_D} - \sqrt{2 \ln k'_D} + \Delta + o(1) \right) \\ &= -2 \left(\ln \frac{k'_D}{\ell_D} \right) \left(\frac{\sqrt{\ln \ell_D}}{\sqrt{\ln k'_D} + \sqrt{\ln \ell_D}} \right) \end{aligned} \quad (18)$$

$$+ \Delta \sqrt{2 \ln \ell_D} + o(\sqrt{\ln \ell_D}) \geq \Delta \sqrt{2 \ln \ell_D} + o(\sqrt{\ln \ell_D}). \quad (19)$$

To check that the last inequality holds, note that Assumption (ii) of the Theorem states that $\liminf \frac{\ell_D}{D} := \eta > 0$, in particular $\ell_D \rightarrow \infty$. On the other hand, since $A_{D,\theta'} \subseteq A_{D,\theta}$, we have $\limsup \frac{k_D}{D} \leq 1 - \eta$, therefore $\limsup \frac{k'_D}{\ell_D} \leq \frac{1-\eta}{\eta} := \gamma$ and finally $\limsup \frac{k'_D}{\ell_D} \leq \max(\gamma, 1)$. Since $\ln \frac{k'_D}{\ell_D} \geq 0$, $\limsup \ln \frac{k'_D}{\ell_D} \leq (\ln \gamma)_+$, and the second factor in (18) is positive and upper bounded by 1, the whole term in (18) is $O(1)$, so that inequality (19) follows.

We deduce that for any $B > 0$ and for D large enough $\frac{b_{\ell_D} - b_{k'_D} + \Delta}{a_{\ell_D}} > B$ holds, and we have

$$\mathbb{P}[\hat{d}_D \in A_{D,\theta'}] \leq \mathbb{P}\left[\left(\frac{\max_{d \leq k'_D} \zeta'_d - b_{k'_D}}{a_{k'_D}}\right) - \frac{\max_{d \leq \ell_D} \zeta_d - b_{\ell_D}}{a_{\ell_D}} > B\right].$$

By similar arguments as in the proof of Theorem 2, the latter upper bound converges to $\mathbb{P}[G - G' > B]$, where G, G' are two independent Gumbel random variables. As B is arbitrary we come to the announced conclusion. \square

References

- D. H. Alexander and K. Lange. Stability selection for genome-wide association. *Genet. Epidemiol.*, 35(7):722–728, 2011.
S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

- F. Bach. Bolasso: model consistent Lasso estimation through the bootstrap. In *Proc. 25th intern. conf. on Machine learning (ICML)*, pages 33–40. ACM, 2008.
- A. Beinrucker, U. Dogan, and G. Blanchard. Early stopping for mutual information based feature selection. In *Proc. 21st Intern. Conf. on Pattern Recognition (ICPR)*, pages 975–978, 2012a.
- A. Beinrucker, U. Dogan, and G. Blanchard. A simple extension of stability feature selection. In *Pattern Recognition*, volume 7476 of *Lecture Notes in Computer Science*, pages 256–265. Springer, 2012b.
- J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res.*, 3:1229–1243, 2003.
- L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.
- P. Bühlmann and B. Yu. Analyzing bagging. *Ann. Statist.*, 30(4):927–961, 2002.
- P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: clustering and sparse estimation. *J. Statist. Plann. Inference*, 143(11):1835–1858, 2013.
- T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, second edition, 2006.
- B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.
- P. Embrechts. *Modelling extremal events: for insurance and finance*, volume 33 of *Stochastic modelling and applied probability*. Springer, 1997.
- G. Escudero, L. Marquez, and G. Rigau. Boosting applied to word sense disambiguation. *Proc. Eur. Conf. on Machine Learning (ECML)*, pages 129–141, 2000.
- F. Fleuret. Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.*, 5:1531–1555, 2004.
- J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.
- I. Guyon. *Feature extraction: foundations and applications*, volume 207. Springer, 2006.
- T. Hastie and B. Efron. *LARS: Least Angle Regression, Lasso and Forward Stagewise*, 2012. URL <http://CRAN.R-project.org/package=lars>. R package version 1.1.
- A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1):145, 2012.
- Q. He and D.-Y. Lin. A variable selection method for genome-wide association studies. *Bioinformatics*, 27(1):1–8, 2011.
- Z. He and W. Yu. Stable feature selection for biomarker discovery. *Comput. Biol. Chem.*, 34(4):215–225, 2010.
- G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
- M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer, 1983.
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.
- K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.
- MASH Consortium. The MASH project. <http://www.mash-project.eu>, 2012. [Online; accessed 19-March-2013].
- N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 72(4):417–473, 2010.
- N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.
- D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer Series in Statistics. Springer, 1999.
- W. Sauerbrei and M. Schumacher. A bootstrap resampling procedure for model building: application to the Cox regression model. *Stat. Med.*, 11(16):2093–2109, 1992.
- R. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.*, 37(3):297–336, 1999.
- R. D. Shah and R. J. Samworth. Variable selection with error control: another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(1):55–80, 2013.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.
- S. Wang, B. Nan, S. Rosset, and J. Zhu. Random Lasso. *Ann. Appl. Stat.*, 5(1):468–485, 2011.